DOCUMENT RESUME

ED 186 474 TM 800 179

AUTHOR Nugent, William A.; Laabs, Gerald J.
TITLE Performance Test Objectivity: Comparison of Interrater Reliabilities of Three Observation Formats.
NOTE 9p.; Paper presented at the Annual Meeting of the Military Testing Association (20th, Oklahoma City, OK, October 30-November 3, 1978).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adults; *Behavior Rating Scales; Competence; *Job Performance; Job Skills; *Military Personnel; Military Training; Reliability; *Student Evaluation; Validity; Videotape Recordings; *Work Sample Tests
IDENTIFIERS *Evaluation Problems; *Interrater Reliability; Navy

ABSTRACT

Comparisons were made of three different procedures for rating job performance and the validity of each procedure was estimated. All three procedures were based on observation of videotaped presentations of 12 electrical measurement problems, 8 of which had errors associated with them. The three procedures involved the use of three different evaluation forms: structured, semi-structured, and unstructured. The unstructured and semi-structured forms were modeled after existing Navy forms; but the structured form, requiring the evaluation of a series of explicit procedural steps, was specially designed for this study. Of the 78 raters tested, 42 were considered proficient and 36 were classified as low skill proficient. The percent agreement of raters using the structured rating form with correct job performance was nearly perfect (97 percent), but the agreement for the semi-structured form was only 81 percent and for the unstructured form was 77 percent. The percent of rating errors of omission and commission was also observed as were differences between high skilled and low skilled raters. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

PERFORMANCE TEST OBJECTIVITY: COMPARISON OF INTERRATER RELIABILITIES OF THREE OBSERVATION FORMATS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY W. Nugent TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

William A. Nugent and Gerald J. Laabs

Navy Personnel Research and Development Center
San Diego, California 92152

TM800 179 ED186474

INTRODUCTION

The current methods for evaluating performance in the Navy consist of a variety of techniques that often are not adequately assessed in terms of validity, reliability, or objectivity prior to their use as measures of job performance. While some of these procedures may yield useable information related to job performance, the accuracy of that information may be limited. For example, a portion of the performance evaluations conducted in the Navy are based upon a rater's judgment concerning observed performance. It is typically assumed that when such "hands-on" performance tests are conducted, the tests and resultant data are valid and reliable. However, if any ambiguity exists in terms of the performance steps to be observed and evaluated; reduced agreement among all the various raters is sure to result, which seriously affects the validity and reliability of the evaluation procedure. The reduced agreement that occurs among raters in this type of evaluative situation stems primarily from a lack of test objectivity.

Objectivity in performance testing refers to the consistency with which raters make their judgments. One of the important variables that directly affects test objectivity is that of test format. Without specific guidelines on what steps or processes to observe, a rater is forced to make subjective judgments that are based on personal standards and prejudices. Raters should not be expected to evaluate steps they cannot see, such as those involved in evaluating a mental process, and each step should be clearly stated. When several ongoing processes are observed and evaluated as a single step, or there is ambiguity as to what constitutes a performance step, it becomes difficult to obtain consistent ratings across raters. On the other hand, the more structured a test format is, the more the raters should agree on completion of steps in a problem.

Another important variable that may interact with test objectivity is the expertise of the rater. The degree of experience that raters have with a particular piece of equipment will influence their judgments of how others use it. Within the context of the Navy's Personnel Qualification Standards (PQS) program, for example, job performance evaluations are conducted by senior supervisory personnel, or by job incumbents that have successfully passed the section to be evaluated. Unfortunately, it is assumed that when raters are qualified in this manner, questions concerning the objectivity of the hands-on performance test are not relevant.

Purpose

The primary purpose of this study was to determine the amount of interrater agreement and reliability of ratings obtained when three structurally different rating formats were used to evaluate the same behavior. In addition, this study examined the relationship between a rater's ability to accurately evaluate the performance of others as a function of his own skill proficiency within a given task area.

METHOD

Stimulus Materials

The appropriate method for determining the consistency of raters' judgments is to hold constant the behavior to be observed and evaluated. One way behavior can be held constant is by videotaping the test behavior so that any variation in evaluation would be due to rater or format differences and not due to test performance differences. Therefore, a videotape was produced in which Navy employees performed four electrical measurements: negative DC voltage measurement, positive DC voltage measurement, and two resistance measures. These measurements were performed using a Simpson Model 260 volt-ohm-meter (VOM) and a Hydrotronics Test Signal Generator. The latter device was specifically designed to provide electronic signals for a previous research project on the use of test equipment (Laabs, Panell, & Pickering, 1977).

The electrical measurement problems were presented to the rater sample three times in the sequence described above. Each type of electrical measurement was performed correctly on only one of the three showings. On the two remaining presentations, the measurements were associated with errors that varied in magnitude. Thus, stimulus materials consisted of 12 videotaped segments in which the four types of electrical measurements were presented in sequential order, while the correct and incorrect performances were presented randomly.

Of the eight videotaped segments that had errors associated with them, only six gave incorrect meter readings. Ratings of these six tape segments were compared to those of the four segments performed correctly. The criterion for the assignment of a pass or fail judgment for each rating form was made on the basis of the meter readings obtained in these ten problems only. The two remaining videotaped segments contained only minor procedural errors and were not included in the present analysis.

The format of the videotaped segments was standardized. Each segment began with a narration of the electrical measurement problem to be performed. Next, the videotape segment showed the steps the examinee used to solve the measurement problem. The videotaped segments ended with the examinee's statement that the problem had been completed, and the examinee's report of the final VOM reading obtained.

Rater Evaluation Forms

One of three different evaluation forms were used by the participants to rate the videotaped performances of the electrical measurements. These forms consisted of a structured, semi-structured, and unstructured format.

The unstructured rating format was modeled after a part of the Personnel Qualifications Standards program. This form required the rater to evaluate overall performance, marking a pass or fail for each measurement problem and recording the errors detected. No structured step-by-step procedures were provided to make the evaluations, nor were any criteria specified for a passing performance on any problem.

The semi-structured rating form is similar to forms the Navy Personnel Research and Development Center has developed and used in the past. The form was adapted from a portion of a performance testing program associated with a self-paced test equipment course that is currently administered at the Submarine Training Center, Charleston, South Carolina. This method required the rater to evaluate the videotaped segments against a number of structured areas of performance, assigning a predetermined weighted value to each area. An area of performance often involved more than one procedural step. When the performance was completed, the rater summed the individual point values assigned to the performance areas to determine whether criterion for passing i.e., 7.5 points out of 10 (or 75% correct) had been met.

The structured rating form was developed specifically for this study. It required the rater to evaluate the videotape segments against a series of procedural steps, each consisting of a single behavior. In addition, this form required that each step be performed in the correct sequence to receive a passing score. The VOM equipment face was reproduced on the form so that the position of control settings, the location of lead connections, and the final meter reading obtained could be easily noted or marked on the response form.

To develop the structured rating format, a preliminary version was presented to 12 Sonar Technician Class "A" School instructors from the Fleet Anti-Submarine Warfare School, San Diego, and they were asked to indicate each step of the procedure that was mandatory to achieve a passing performance for each problem. The final version of the structured rating form consisted only of those steps that 85% of the instructors considered essential for passing. Furthermore, there was general agreement among the instructors on the sequential order for the completion of the steps that were retained for each of the four measurement problems.

Rater Expertise

The second independent variable studied was rater expertise. Expertise level was determined by the score the rater obtained on a VOM proficiency test. This test consisted of the same four types of electrical measurement problems that the raters were asked to evaluate during the videotaped presentations. The VOM proficiency test also used the identical equipments as those used in the production of the videotaped segments.

Two proficiency level categories were established for the rater expertise variable: raters who passed two or more problems out of four were considered to be high skill proficient; whereas raters who failed to pass at least two of the four problems were considered to be low skill proficient. The structured rating form was used by a member of the research staff to evaluate the proficiency level of the raters.

Procedures

Testing was conducted in an experimental laboratory at the Navy Personnel Research and Development Center, San Diego. One half the rater sample received the VOM proficiency test prior to the videotaped presentations; half after viewing the videotape. In both conditions, raters were tested individually on the VOM proficiency test and each rater was assigned to use one of the three rating forms on a random basis.

Raters conducted their evaluations of the videotaped segments in groups of two or three at individual television monitor carrels so that one rater's judgment would not influence another. Prior to evaluating the videotape segments, raters were given a practice session to become familiar with the composition of the videotaped presentations as well as the rating format they had been assigned. The raters viewed each segment, consisting of a single electrical measurement problem, only once. Following each segment presentation, raters were given a 30-second time period to complete entries to their rating forms. The forms were collected when the raters had completed their evaluation of the final videotaped segment.

As discussed previously, the three rating formats differed from one another with respect to the process by which performance steps were observed and evaluated for each videotaped segment. However, the three forms were comparable in that that they provided raters with a means of judging the overall product (i.e., assigning a passing or failing score for each electrical measurement problem). Consequently, the criterion by which the performance of the raters was measured involved comparison of the rater's dichotomous pass/fail responses for each segment to the predetermined standard for the 10 videotaped presentations that were analyzed.

Sample

A total of 15 instructors and 63 students from the Anti-Submarine Warfare School, San Diego, participated in the study. The students in the study were either designated Sonar Technicians or were undergoing Class "A" School training in that rating.

Of the 78 raters tested; 28, 26, and 24 raters were assigned on a random basis to the structured, semi-structured, and unstructured format, respectively. On the basis of the VOM proficiency test, 16 of the raters who used the structured format were classified as high skill proficient and 12 as low skill proficient. Of the raters who used the semi-structured format, 14 were classified as high skill proficient and 12 as low skill proficient. Finally, 12 of the raters who used the unstructured format were classified as high skill proficient and 12 were classified as low skill proficient.

RESULTS

Proficiency Test/Rating Form Presentation Order

No differences were found in terms of criterion agreement with the videotaped presentations as a function of whether the VOM proficiency test was given before or after viewing the videotape. Significant differences also failed to appear in terms of correct performances on the VOM proficiency test as a function of whether the videotaped presentations were shown before or

after the VOM proficiency test. Therefore, for all remaining analyses, raters were collapsed across presentation order.

Interrater Reliability

An estimate of interrater reliability was calculated for each form through application of the dichotomous pass/fail responses to an analysis of variance technique that yields an intra-class correlation (Winer, 1971, p. 283). It was found that raters who used the structured rating form showed the highest interrater reliability with a coefficient of reliability of .996. The reliability coefficients for the semi-structured and unstructured formats were .973 and .808, respectively. These coefficient differences were tested by a chi square analysis (Snedecor & Cochran, 1967, p. 286) and were found to be statistically significant ($\chi^2 = 42.4$, df = 2, $p < .001$). Although the structured rating form had the highest coefficient of interrater reliability, the semi-structured and unstructured forms appear to have acceptable levels of interrater reliability in terms of evaluating overall performance on the videotaped segments.

No significant differences were found in interrater reliability values within each rating form as a function of rater skill proficiency.

Criterion Agreement

Table 1 provides a summary of the mean percent agreement with the predetermined pass/fail criterion across the three rating formats. The table shows that the use of the structured rating format resulted in the highest average percent of criterion agreement, while the use of the semi-structured and unstructured formats resulted in progressively less average agreement.

Table 1

Mean Percent Agreement with the Pass/Fail Criterion

Across Three Formats

	Rating Format		
	Structured	Semi-Structured	Unstructured
M	97.1	80.7	76.7
SD	4.6	12.0	14.3

Individual rater percentage values across all three rating forms were converted to standard scores, and an analysis of variance was performed. The main effect of rating format was found to be statistically significant ($F(2,75) = 26.34$, $p < .001$). A Scheffe post hoc analysis of the mean values revealed that the structured rating format differed significantly from the semi-structured and unstructured formats at the $p < .01$ level. An estimate of the overall strength of association between rating format and criterion agreement was also calculated. The estimate showed that 39 percent of the variance in the dependent variable can be accounted for by the independent variable.

No significant differences were found in the amount of criterion agreement as a function of rater skill proficiency.

Observation Errors on Failed Problems

The above findings clearly indicate that product judgments (i.e., assigning pass/fail scores) are best made using the structured format. However, these data do not fully describe the state of affairs in using the different formats because they do not reflect the errors made in observing the processes or the procedural steps in the electrical measurement problems. For example, the assignment of a failing score that was in agreement with the predetermined criterion could be made for the wrong reason. This might involve an error of omission (failure to identify an incorrect procedural step) coupled with an error of commission (identifying a correctly performed procedural step as incorrect). Although the three formats were, by design, not equivalent in terms of the amount of information related to process judgments, it was felt that a more detailed examination of the errors made when observing the six videotapes of incorrect performances would be useful.

Table 2 shows the average percent of errors of omission for the three formats. For the structured and semi-structured formats, this means that

Table 2

Mean Percent of Errors of Omission for Three Formats

	Rating Format		
	Structured	Semi-Structured	Unstructured
M	7.1	20.2	50.5
SD	8.6	13.3	28.2

an incorrect step was marked as correct or that points were not subtracted for the incorrect step, respectively. For the unstructured format, this means that the error was not written down. Unfortunately, there is no way of determining whether the rater did observe the incorrectly performed step but merely neglected to enter the error on the observation sheet. Thus, the percent of errors of omission for this format might be inflated. Nevertheless, there was a much lower percent of errors of omission associated with the structured format, which supports the findings on criterion agreement across rating formats.

The other error that could occur on the six failure trials is that of commission. For the structured and semi-structured formats, this means a correct step was marked as incorrect or that points were subtracted for a correct step, respectively. For the unstructured format, this means that a correct step was written down as incorrect. Again, there is no way of knowing if other correct steps were observed as incorrect but simply not entered on the observation sheet. Table 3 shows the percent of raters at both skill levels, and within each rating format, that committed at least one error of commission. Inspection of the table shows that skill proficiency of the rater does not appear to make a difference unless the structured format is used to observe the performance. Overall, the structured rating format is associated with the lowest percentage of raters committing errors of commission (46.4%), with the semi-structured and unstructured formats showing much higher percentages of raters committing these errors (92.3% and 95.8%, respectively).

Table 3

Percent of Raters Making Errors of Commission Across Three Formats and Two Skill Categories

	Rating Format		
Skill Category	Structured	Semi-Structured	Unstructured
High	18.5	92.9	91.7
Low	83.8	91.7	100.0

CONCLUSIONS

A drop from almost perfect agreement with the overall pass/fail criterion when raters used the structured rating format to about 77 percent when raters used the unstructured format, demonstrates the importance of providing a list of unambiguous step-by-step procedures to be checked-off when observing hands-on performance. This finding is further reinforced by the fewer errors of omission and commission committed by this group.

It is interesting to note that the relatively poorer showing for the semi-structured and unstructured formats in terms of overall criterion agreement, and errors of omission and commision occurred for both the high skill and low skill proficient groups. This means that being an expert in a given performance area does not necessarily guarantee that all steps in a given job task will be correctly observed and evaluated by raters who used these performance evaluation forms.

The listing of unambigious step-by-step procedures also resulted in high interrater reliability or objectivity for the structured rating format. With less structure in the rating format, there was less objectivity in observing and evaluating both passing and failing performances. In addition, the level of rater skill proficiency became more important on the structured rating form when errors of commission were examined. Significantly fewer raters in the high skill proficient group made errors of commission than in the low skill proficient group ($t = 3.38$, $df = 25$, $p < .01$).

This finding suggests that high skill proficient raters are more apt to accurately observe and evaluate the process by which the electrical measurements were performed. The failure to achieve significant differences between high and low skill proficient raters with respect to commission errors on the two remaining formats may be attributed to a lack of specificity in the performance steps to be observed and evaluated. Thus, no matter what the format of the observation form to be used, the skill proficiency of a rater should probably not be ignored.

The unstructured and semi-structured formats are presently in use in the Navy to evaluate hands-on job performance. It is clear that if these rating formats are replaced by more structured rating forms; more reliable, valid, and objective measurements of hands-on job performance would result.

REFERENCES

Laabs, G. J., Panell, R. C., & Pickering, E. J. A personnel readiness training program. Maintenance of the Missile Test and Readiness Equipment (MTRE MK 7 MOD 2) (NPRDC Tech. Rep. 77-19). San Diego: Navy Personnel Research and Development Center, March 1977. (AD-A037 546)

Snedecor, G. W., & Cochran, W. G. Statistical methods (6th Ed). Ames: Iowa State University Press, 1967.

Winer, B. J. Statistical principles in experimental design (2nd Ed.). New York: McGraw-Hill, 1971.